**Credit Card Clients: Project report**

GitHub repository URL: https://github.com/ann-marie1983/Credit-card-data-analysis

**Introduction**

I selected the dataset ‘Credit card clients: Predict churning clients’ which I found on Kaggle. What struck me about this particular dataset is the problem statement that came with it. Its real-life scenario that I can relate to from a business perspective. If any businesses could predict client attrition and then target those clients prior to losing their custom, it could be more beneficial than spending large sums of money trying to appeal to new clients

**Objective of data analysis**

My goal is to look for commonality in the type of clients that leave. I need to therefore look at aspects of the ‘churn’ client that show commonality.

**Description of Dataset**

The dataset selected ‘Bank Churners’ was downloaded from Kaggle. The data set is tabular, consisting of 23 columns and 10,127 rows. There are 23 headings detailing personal attributes such as client number, age, Gender, education, marital status and income etc as well as client attributes such as card category, months on the books, inactive months, credit limit etc. We need to see what personal attributes influence lead to client attrition and if the client attributes are part of the issue.

Source of data: <https://leaps.analyttica.com/home>, via [https://www.kaggle.com/sakshigoyal7/credit-card-clients](https://www.kaggle.com/sakshigoyal7/credit-card-customers).)

**Project phases:**

**Data review and data cleanse**

Firstly, I imported Panda and NumPy libraries to assist me in my data review and data analysis phases of this project. The dataset I chose is tabular but mostly consists of numerical data so both Pandas and NumPy are essential.

I then imported by csv file to a Pandas data frame in Pycharm. I checked if the client number could be used as the primary key for this dataset. I did this by comparing the outcome of ‘len(data.CLIENTNUM.unique())’ to the number of entries output by ‘(data.info()’.

On initial review of the data, I chose to use the function (data.info() instead of separate functions (data.head()), (data.shape), (data.dtypes) and (data.columns) as this function pulls in these details as well as checking for null values in the data. It also identified that in the data set the data types are 7 floats, 10 integers and only 6 objects.

I checked the data for duplicates and there are none, if just a few duplicates had appeared I would have removed them.

Though there were no missing values I did include a default code ‘cleaned data’ had the column ‘Months on books’ contained null values. I consider this an important attribute when evaluating the attrition of client and so paid special attention to it from the perspective of missing value, keeping in mind this code could be run on the same updated dataset in the future, So I made the decision to replace these values with the median value. The reason I did this, was had a null value appeared in this important attribute I would not like to delete this client as the dataset is comparatively small. When I checked the mean, median and mode values they came back to a very similar figure so I opted to use the median as a replacement knowing it would not skew the data. Of course, if there was a significate number of missing values for ‘Months on the books’, I would need to consider another approach. We looked for columns with null values knowing there are none I then looked to drop the last two columns of this dataset that I felt would not assist me in my analysis.

Before I began my analysis, I wanted to get a glimpse of a client’s personal data using indexing and slicing together. So, I could get a feel for a typical client. I then checked the credit card types to see how many types were available.

**Data analysis**

Overview

* They lost 1,627 clients which is 16% of their total client base.
* I created a list, I viewed a small sample to check quickly was there any correlation between the card category and the income category of the employee. List has more functionality than the tuple, so I preferred to use a list in this instance.
* I did notice the Blue card was very popular. Checking the count for each Card category I realised the vast majority of clients have the Blue card. I also gave an example of how this data could be used as a data frame and combined with set data such as a credit card limit per card type.
* By counting the income categories I could now see that most credit card clients earn less than $40K per year. Also, there are a large number of clients whose income is ‘Unknown’.
* I examined the age profile of a client, the mean age of a client which is 46. I checked how many fell into age categories ‘great than or equal’ to 30, 46, 56 and 66 to check for any concentration in age at any particular age group. As expected over half were 46 and above but very few clients were 66 and above.
* I wanted to see what kind of transaction activity these clients had so I created a feature to get a better understanding. The transaction activity per month is low in the small sample I viewed.

Examining Attrite clients

* I checked how many months on average did attrite clients leave in. Then queried how long existing clients were on the books. I noticed that Attrite clients tend to leave after 36 months. I also noticed that the existing clients have been clients for less than 36 months. This is concerning and shows the need to prevent attrition may be quite urgent.
* I looked at the client’s interactions with the company based on attrite versus existing clients and noticed attrite clients had a lower relationship count on average. This would indicate that this attribute is something that must be monitored.
* During my data review I noticed the majority of data types in my dataset were Integers and floats, so I grouped all by Attrition flag and examined their inactivity, contacts in the last 12 months, credit limits, Total revolving balance, total transaction amounts and their utilisation ratio. I used grouping here as it allowed me to focus on the most important attribute the ‘Attrition\_Flag’.
  + Their months inactive mean value for attrite clients is higher as expected however the contacts count in the last 12 months is higher.
  + Next, I checked the credit limit and noticed the credit limit for existing clients tend to be higher. This would raise the question are clients leaving because of their credit limit?
  + The total revolving balance is less for attrite clients indicating possible less usage before they left.
  + The total transaction amounts for existing clients are higher
  + The mean average utilization rate confirms that the attrite clients used their card less

Overall, we can surmise attrite clients used their card less, had lower transaction amounts and a lower credit limit. I would consider if their need to borrow was les and so this is why they are no longer clients.

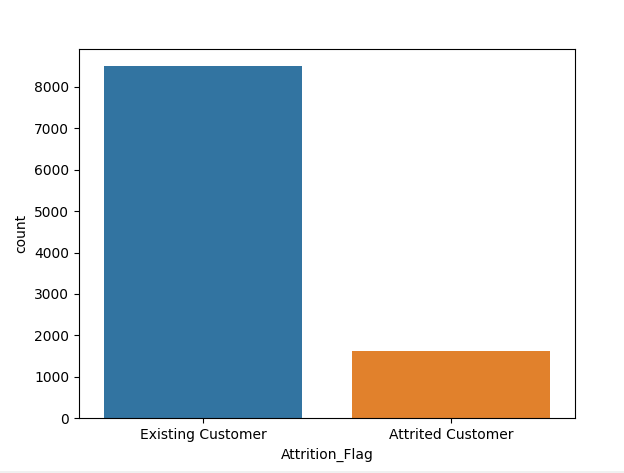
* When reviewing the string data, I also grouped them by Attrition flag but used value counts. I assessed the client’s marital status, Education level, card category and income category. Credit card clients tend to be married or single, attrite clients that are single are disproportionate to existing clients that are single. I also noticed clients earning Less than $40K tend to be disproportionally attrite.

**Insights from my Data analysis on plots and graphs**

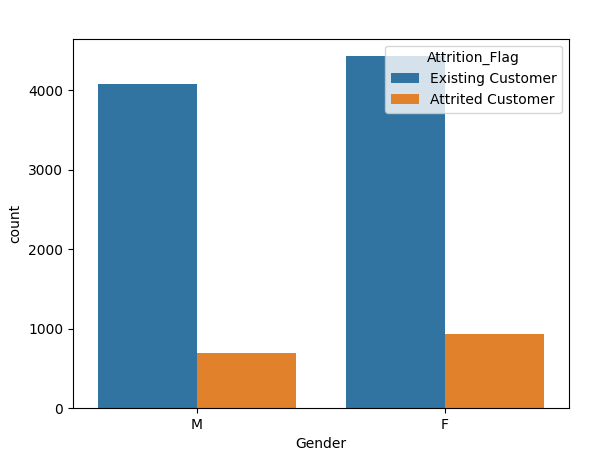
Before creating any visualisation using information from my data analysis, I imported the relevant libraries Matplotlib.pyplot and Seaborn. I concentrated client’s personal attributes to understand a typical and Attrite client. Then I concentrated on credit card attributes, client’s usage of their credit card and attrition, then finally the client relationship and company contacts with the client.

Client Attributes

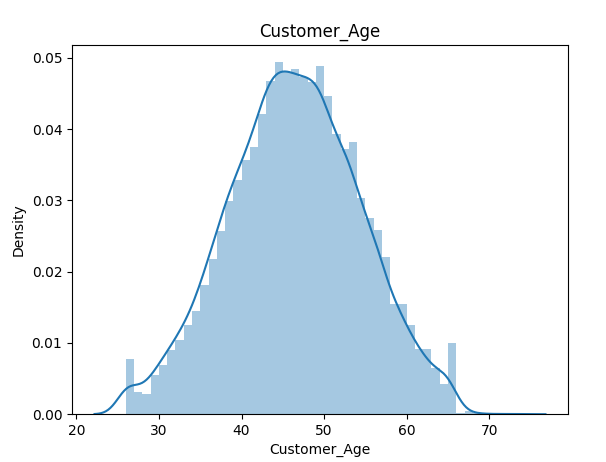
From the below bar chart, we can see the number of Attrite clients to existing clients, there are very few Attrite clients, approximately 16% of the Total client base.



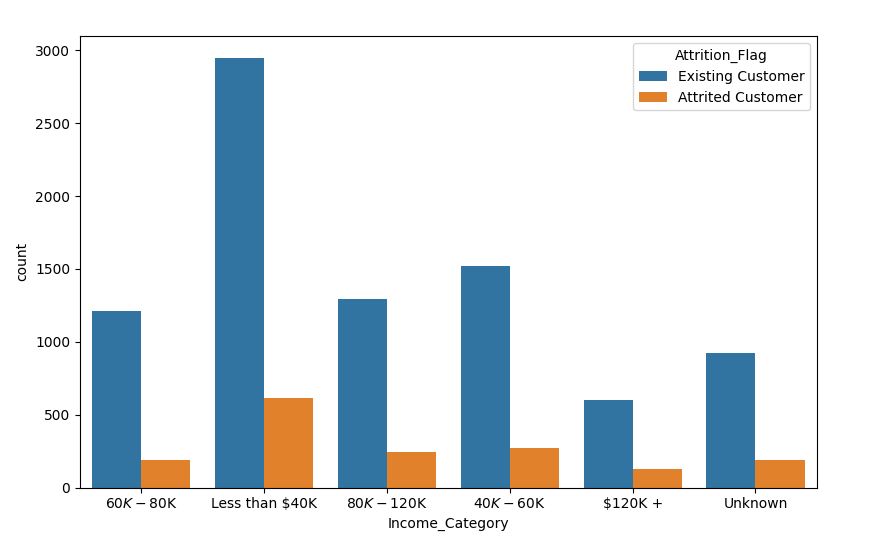
Next, I plotted the client’s gender in relation to attrition. We can see from below that there are slightly more female clients but both Male and Female clients are similarly likely to churn.



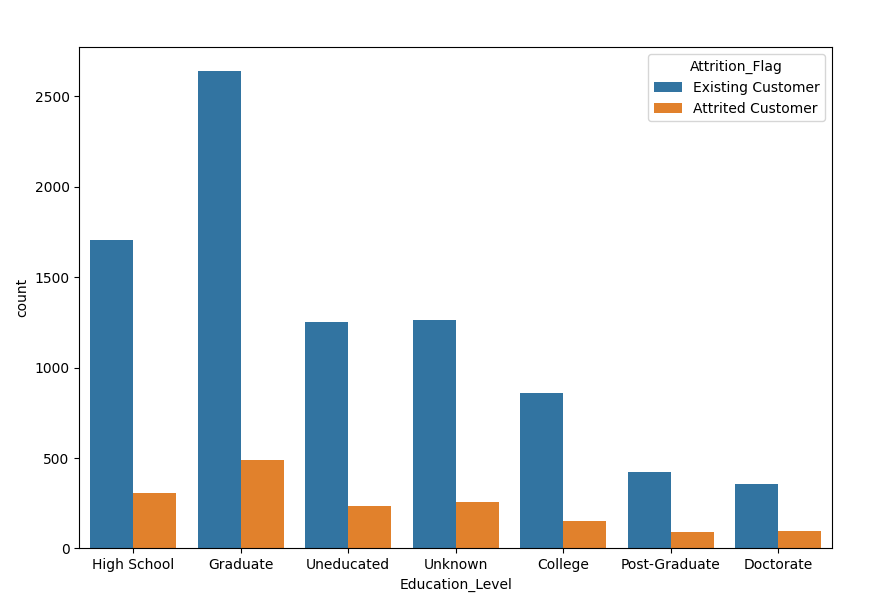
We already know the mean and median age of existing clients versus attrite clients is very similar, so I used this graph to understand the overall age profile of clients. The vast majority of clients are between 35 and 60 years of age.



Another key attribute of the client data was income category, we can see the vast majority of clients earn less than $40K per annum. Visually it looks like this category of clients contains a disproportionate number of attrite clients. This is something I will look into further.

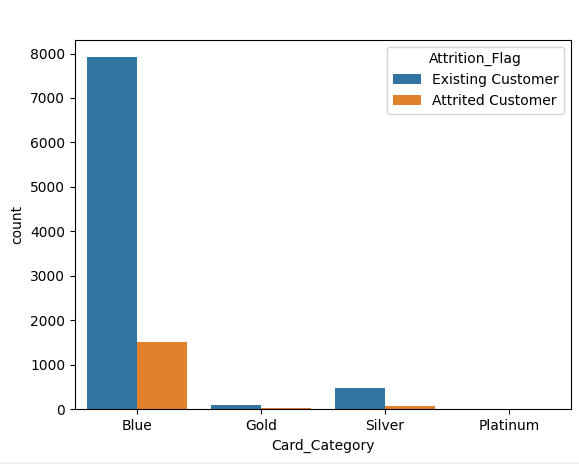


Most clients are Graduates, we can see from this plot that most customers are fall into Educational levels ‘Uneducated’, ‘High school’ and ‘Graduates’. The Unknow Education level of this attribute is large and so this attribute though useful as a general indication of Attrition may not be usable, as I would not want to replace the ‘unknown’ value.

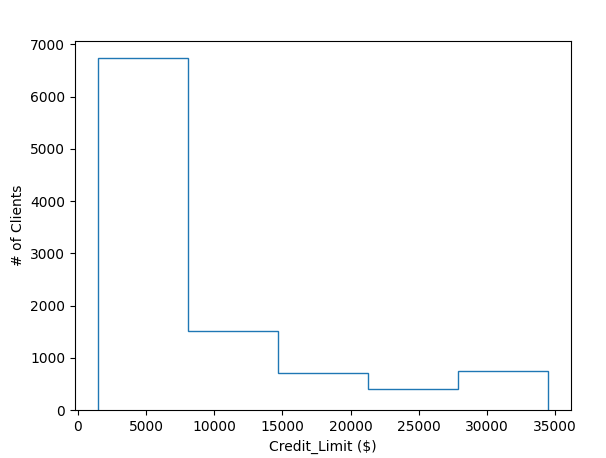


Credit card Attributes

The majority of clients are blue card holders, attrite clients do not seem to be a particular card holder type. This would indicate that the card type is not a useful attribute when trying to profile a typical client that may leave.

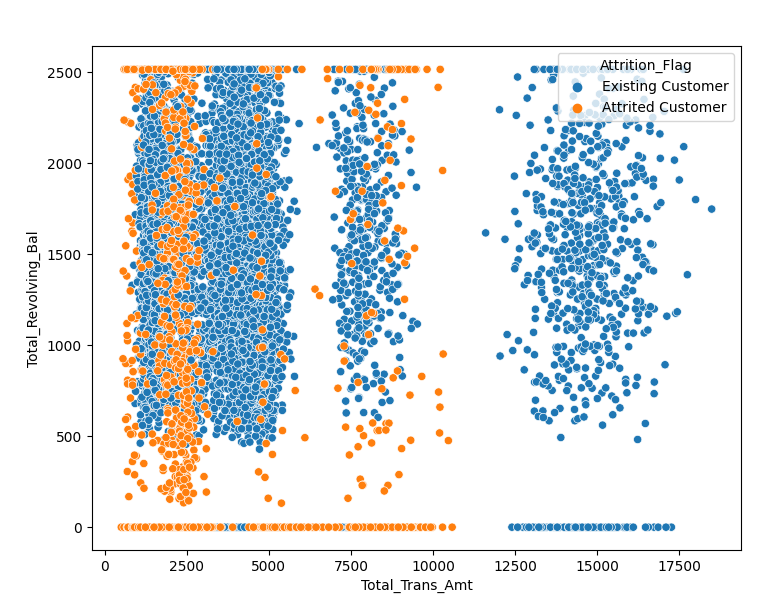


To get a better understanding of their clients both Attrite and Existing I wanted to understand the clients credit limit. Most clients have a credit limit below $10K. This makes sense as most clients are low earners and would borrow in line with their income.

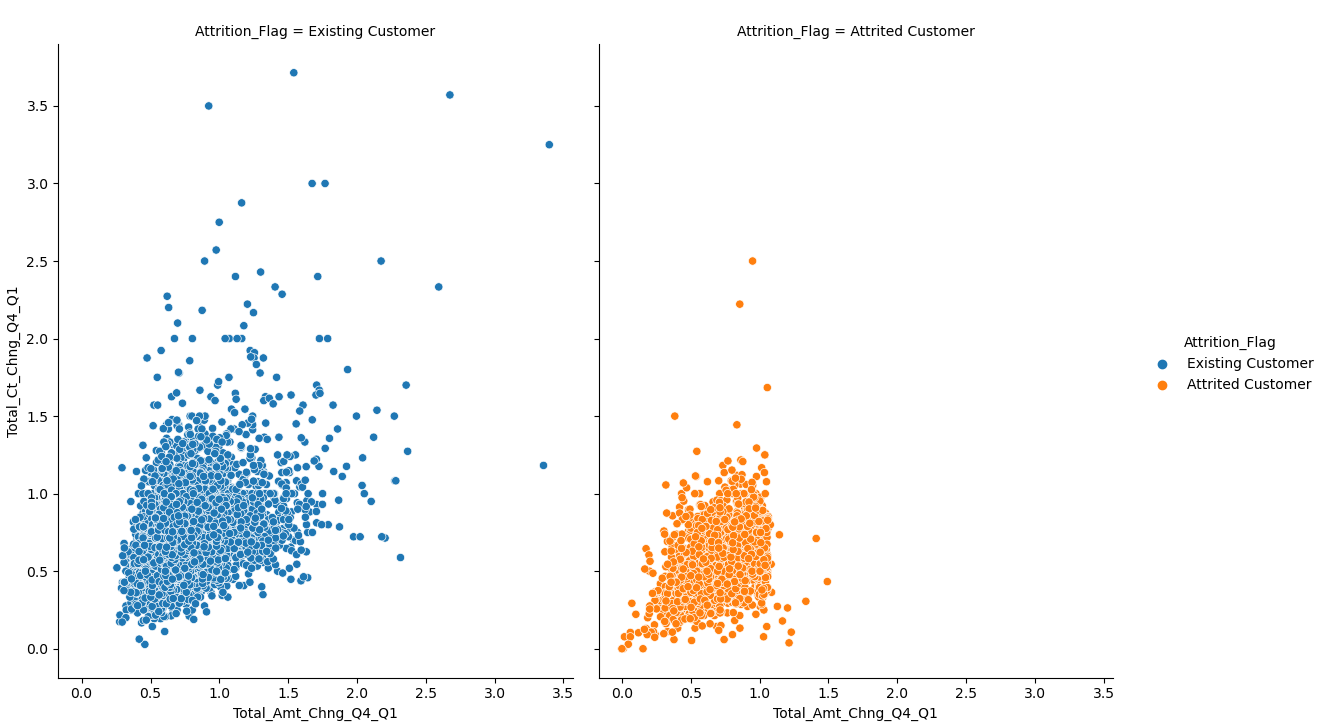


Credit card usage

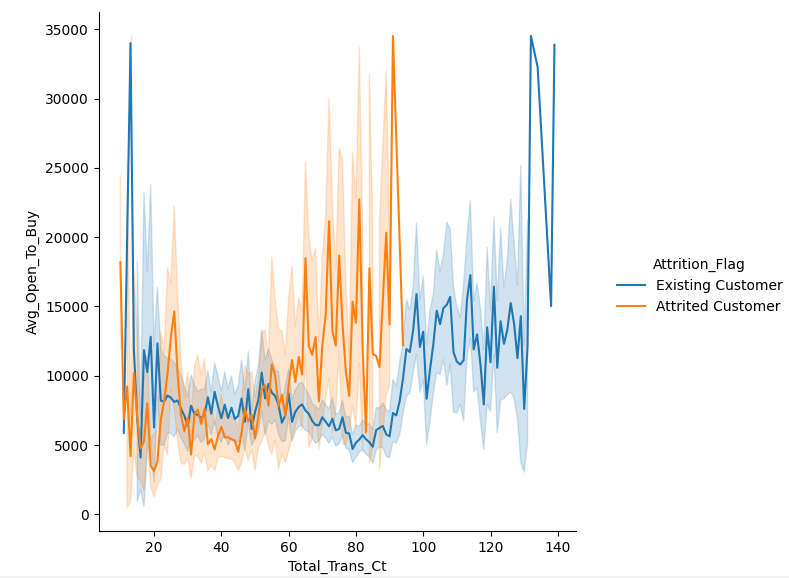
Now that I understand the clients better, I wanted to look at their usage of their credit cards. The revolving balance of attrite clients is quite evenly spread with a lot of balances remaining below that of an existing client. Also, the total transaction amounts are much less for attrite clients. This would indicate less usage of their credit card.



Y = Change in Transaction Count, versus X = Change in Transaction Amount, this graph shows existing clients had more transactions and they were for greater amounts. Attrite client’s usage was less during their relationship with the credit card company.



I then checked what amount of credit was available (‘Open to buy’) to existing and attrite clients versus the number of transactions. Attrite customers used less transactions but higher amounts getting closer to their credit limit. Existing clients seem to use their credit in a more consistent way, many transactions take them closer to their limit. Do Attrite clients get a credit card for a particular purpose and once paid off no longer require a credit card or are they going to a competitor that will allow them to have a higher credit limit. This data would need to be gathered.



Interestingly we can a large group of Attrite clients with zero ‘No. of Contacts in the last 12 months’ had a relationship length between 25 to 38 months with a mean of 36 months. We also have a group of attrite clients that have been in contact 6 times which is very high (no existing clients have been in contact 6 times in the last 12 months which would indicate a possible issue. The contacts count seems to be a variable worth investigating.

